High-resolution mapping of *cis*-regulatory variation in budding yeast

Ryosuke Kita[a], Sandeep Venkataram[a], Yiqi Zhou[a], and Hunter B. Fraser[a,1]

[a]Department of Biology, Stanford University, Stanford, CA 94305

Edited by Jasper Rine, University of California, Berkeley, CA, and approved November 3, 2017 (received for review October 4, 2017)

Genetic variants affecting gene-expression levels are a major source of phenotypic variation. The approximate locations of these variants can be mapped as expression quantitative trait loci (eQTLs); however, a major limitation of eQTLs is their low resolution, which precludes investigation of the causal variants and their molecular mechanisms. Here we report RNA-seq and full genome sequences for 85 diverse isolates of the yeast *Saccharomyces cerevisiae*—including wild, domesticated, and human clinical strains—which allowed us to perform eQTL mapping with 50-fold higher resolution than previously possible. In addition to variants in promoters, we uncovered an important role for variants in 3′UTRs, especially those affecting binding of the PUF family of RNA-binding proteins. The eQTLs are predominantly under negative selection, particularly those affecting essential genes and conserved genes. However, applying the sign test for lineage-specific selection revealed the polygenic up-regulation of dozens of biofilm suppressor genes in strains isolated from human patients, consistent with the key role of biofilms in fungal pathogenicity. In addition, a single variant in the promoter of a biofilm suppressor, *NIT3*, showed the strongest genome-wide association with clinical origin. Altogether, our results demonstrate the power of high-resolution eQTL mapping in understanding the molecular mechanisms of regulatory variation, as well as the natural selection acting on this variation that drives adaptation to environments, ranging from laboratories to vineyards to the human body.

yeast | eQTL | gene expression | population | evolution

Genome-wide association studies (GWAS) have identified thousands of associations between genetic variants and phenotypes in a wide range of species. As more of these associations are identified, there is a concomitantly increasing need to answer the question of how these variants shape particular phenotypes. A common mechanism for these phenotype-altering variants is via changes in gene expression (1–3). Indeed, human disease-associated variants are highly enriched for regulatory functions, and genetic variants associated with gene expression can implicate causal genes (4–6). In addition, gene-expression regulation has been found to be the predominant target of positive selection in recent human evolution (7, 8). These findings all support the longstanding hypothesis that *cis*-regulatory variants are a critical component in the evolution of complex traits.

The budding yeast *Saccharomyces cerevisiae* is a key model organism for investigating how genetic variants influence gene expression. Genetic variants or loci associated with a gene's mRNA abundance are known as expression quantitative trait loci (eQTLs). Recombinant lines of two *S. cerevisiae* strains were utilized for the first genome-wide eQTL mapping (9). This work identified widespread "local" eQTLs located very close to the regulated gene, which are predominantly caused by *cis*-acting variants (10), as well as several transacting "hotspots," where a single locus controls the expression of many genes (9). Additional studies on the same genetic cross identified condition-specific eQTLs (11), genetic interactions between eQTLs resulting in nonadditive effects (12), and widespread adaptive evolution of gene expression (13).

Saccharomyces hybrids have also proven to be a valuable resource to study *cis*-regulatory variation, since allele-specific expression (ASE) in a hybrid reflects only *cis*-acting divergence (10, 14), as opposed to *trans*-acting divergence that affects both alleles. Studies of hybrid ASE have revealed interesting cases of pathway-level regulatory divergence (15, 16) and examples of polygenic adaptation affecting traits such as pathogenicity, ergosterol biosynthesis, and toxin resistance (17–19). A key difference between hybrid ASE and eQTL mapping is that ASE reveals the genes affected by *cis*-regulatory variation, but does not indicate the locations of the causal variants as eQTL mapping can.

Despite the utility of eQTL studies in *S. cerevisiae*, they have had several limitations. First, existing eQTLs generally span many kilobases containing dozens of genetic variants, and thus cannot determine the precise location of the causal variant. This is because the mapping was performed with first-generation meiotic segregants with a limited number of recombinations between parental genomes; these recombination breakpoints are needed to map QTLs, since only when the linked alleles are separated by recombination can their effects be distinguished. Second, yeast eQTLs have been mapped from genetic crosses between just a few parental strains, and thus do not sample most of the natural variation across the species. Incorporating a diverse collection of strains to map eQTLs may thus allow not only greater mapping resolution, but also a deeper understanding of species-wide patterns of natural selection on regulatory variation.

Here we address both of these limitations by mapping eQTLs in a diverse set of 85 *S. cerevisiae* strains. By performing eQTL mapping in a wide variety of genetic backgrounds, we minimized linkage disequilibrium between nearby variants and thus mapped eQTLs with high resolution, similar to GWAS in humans and other species (20, 21). In addition, these 85 strains were isolated

Significance

Genetic variants affecting gene-expression levels are a major source of phenotypic variation. Using 85 diverse isolates of *Saccharomyces cerevisiae*, we mapped genetic variants that affect gene expression with 50-fold higher resolution than previously possible. By doing so, we were able to pinpoint likely causal variants and investigate their molecular mechanisms. We found that these genetic variants are generally under negative selection, but also that clinical yeast isolates have undergone positive selection for up-regulation of genes involved in biofilm suppression. Altogether, our results demonstrate the power of high-resolution mapping of genetic variants that affect gene expression, particularly in understanding the molecular mechanisms of regulatory variation and the natural selection acting on this variation.

Author contributions: R.K. and H.B.F. designed research; R.K., S.V., and Y.Z. performed research; R.K. analyzed data; and R.K. and H.B.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All RNA-Seq and DNA-Seq data are deposited in the NCBI Sequence Read Archive (Bioproject PRJNA342356).

[1]To whom correspondence should be addressed. Email: hbfraser@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717421114/-/DCSupplemental.

from a wide range of ecological niches, including clinical isolates, laboratory strains, and vineyard isolates, and thus allowed us to investigate the evolution of gene-expression regulation across these diverse environments.

Results

DNA and RNA-Seq of 85 Yeast Isolates. We obtained 85 *S. cerevisiae* strains from across the world, which included domesticated, wild, and clinical isolates (22) (Fig. 1*A* and *SI Appendix*, Table S1). The domesticated isolates are laboratory strains and vineyard strains; the wild isolates were obtained from wild plants, such as oak trees; and the clinical strains were isolated from sites of infection in human patients. We performed whole-genome sequencing of the 85 isolates, identifying 246,335 nonsingleton variants (205,016 nonsingleton SNPs) (*Methods* and *SI Appendix*, Fig. S1). Sixty-five of these isolates were recently sequenced independently (23), and we observed strong concordance of variant calls for these strains (*Methods* and *SI Appendix*, Fig. S2). Using variants within conserved chromosomal regions, we created a phylogenetic tree (Fig. 1*B* and *SI Appendix*, *Text S1*). The population structure was somewhat associated with geographical origin (e.g., with a tightly grouped set of vineyard isolates from Italy), but there were also clusters of similar ecological niche that spanned multiple continents (*SI Appendix*, Fig. S3), as previously observed (23–25).

To measure gene-expression levels across all 85 isolates, we performed RNA-seq in YPD at 30 °C (*Methods* and *SI Appendix*, Fig. S4). To control for potential read-mapping biases that could cause reads with nonreference alleles to map less well, we masked the reference genome before mapping, converting all single-nucleotide variants identified in our 85 strains to N's. We found no association between mapping rates and divergence from the reference genome, suggesting that mapping bias was not a major issue (*SI Appendix*, Fig. S4). Using both DNA- and RNA-seq data, we also assessed whether any of the 6,572 genes from the reference genome were absent in the genome of each isolate (*SI Appendix*, *Text S2* and Fig. S5). We found a median of 30 genes that were missing across the isolates, located primarily in the subtelomeric regions (*SI Appendix*, Fig. S6). These genes were removed from eQTL mapping analyses, as described below.

Because we identified population structure based on ecological niche in the genomic variants, we asked whether the transcriptomes also exhibit this structure. To test this, we computed the Euclidean distance across all gene expression levels between every pair of isolates, and used these to construct a neighbor-joining tree (Fig. 1*C*). Isolates from the same ecological niche were visibly less clustered together than in the genomic tree, with the large clinical and wild clades split up and interspersed with isolates from other niches. To quantify these relationships, we calculated the first six principal components of the gene-expression profiles (*Methods*). We found only a weak association between the principal components and ecological origin or sequencing batch, suggesting that these are not the primary determinants of these transcriptomes (*SI Appendix*, Figs. S7 and S8). On the other hand, the principal components calculated from the genotypes exhibited strong associations with ecological origin (*SI Appendix*, Fig. S9). The weaker ecological associations with gene-expression principal components may reflect negative selection that has constrained gene expression divergence more than sequence divergence.

High-Resolution eQTL Mapping. To investigate the genetic basis of transcription across these isolates, we mapped the expression level of each gene to variants within 2.5 kb of the transcript boundaries (median of 59 variants analyzed per gene) (*Methods*). We used genome-wide efficient mixed model association (GEMMA) (26) to perform this mapping, as it has been shown to adequately control for population structure in *S. cerevisiae* local eQTL mapping (27). We discovered 1,403 genes with a local eQTL at false-discovery rate (FDR) < 0.05 (Dataset S1). To replicate these eQTLs, we compared them with a previous study that performed RNA-seq on 22 yeast isolates (28). We performed the eQTL analysis using the same method as above, and found significant enrichment of low *P* values and concordant directionality with our eQTLs (*SI Appendix*, *Text S3* and Fig. S10).

Because we mapped eQTLs using a large number of isolates with high genetic diversity, we hypothesized that our mapping resolution might be higher than previous eQTL analyses in *S. cerevisiae*. To test this, we compared the resolution of the eQTLs from this study with the resolution from a previous analysis of 112 segregants from a genetic cross (10). QTL widths are typically reported as a 1-LOD (logarithm-of-odds) or 2-LOD support interval (defined as the distance between two genetic markers: the first marker to the left of the most significant marker in a QTL to have a LOD score at least 1 or 2 lower than the most significant marker, and the equivalent marker to the right). We observed a 49.7-fold higher resolution in our analysis with a median 1-LOD interval of 1,210 bp (containing a median of 14 variants; 462 intervals, or 33%, had exactly one variant), compared with the median of 60,230 bp in the segregant analysis ($P < 10^{-15}$) (*Methods* and Fig. 2*A*). We replicated this finding with 2-LOD support intervals from the same study, and 1.5-LOD support intervals from another yeast eQTL study (11) (*SI Appendix*, Fig. S12*A*).

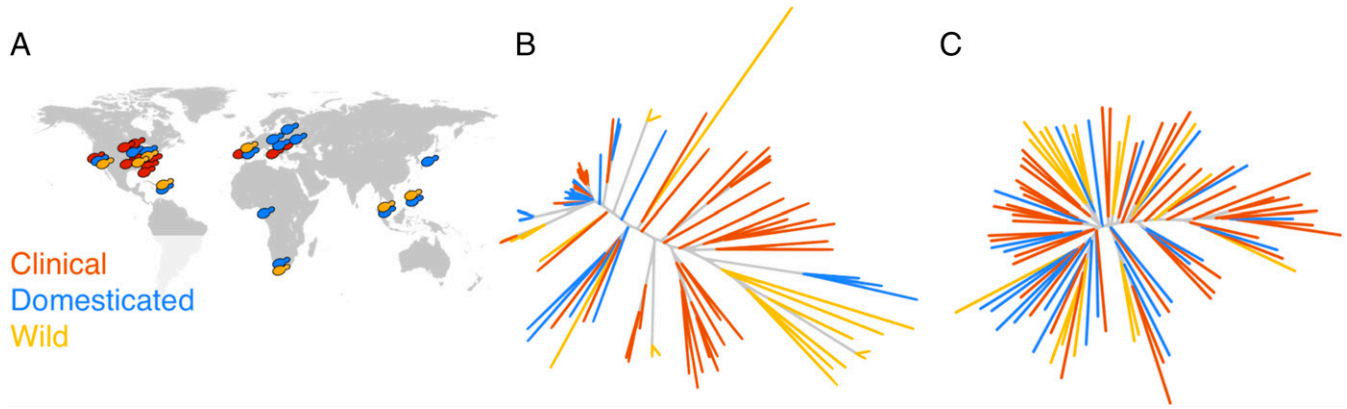

Fig. 1. Origins and relationships among 85 isolates. (*A*) The isolation locations of the strains analyzed in this study. Colors represent the ecological niche of origin. (*B*) Neighbor-joining tree created using a 218-kb sequence comprised of conserved regions across the genome. Colors label the origin of the strain. (*C*) Neighbor-joining tree created using the Euclidean distance of the normalized expression across all genes.

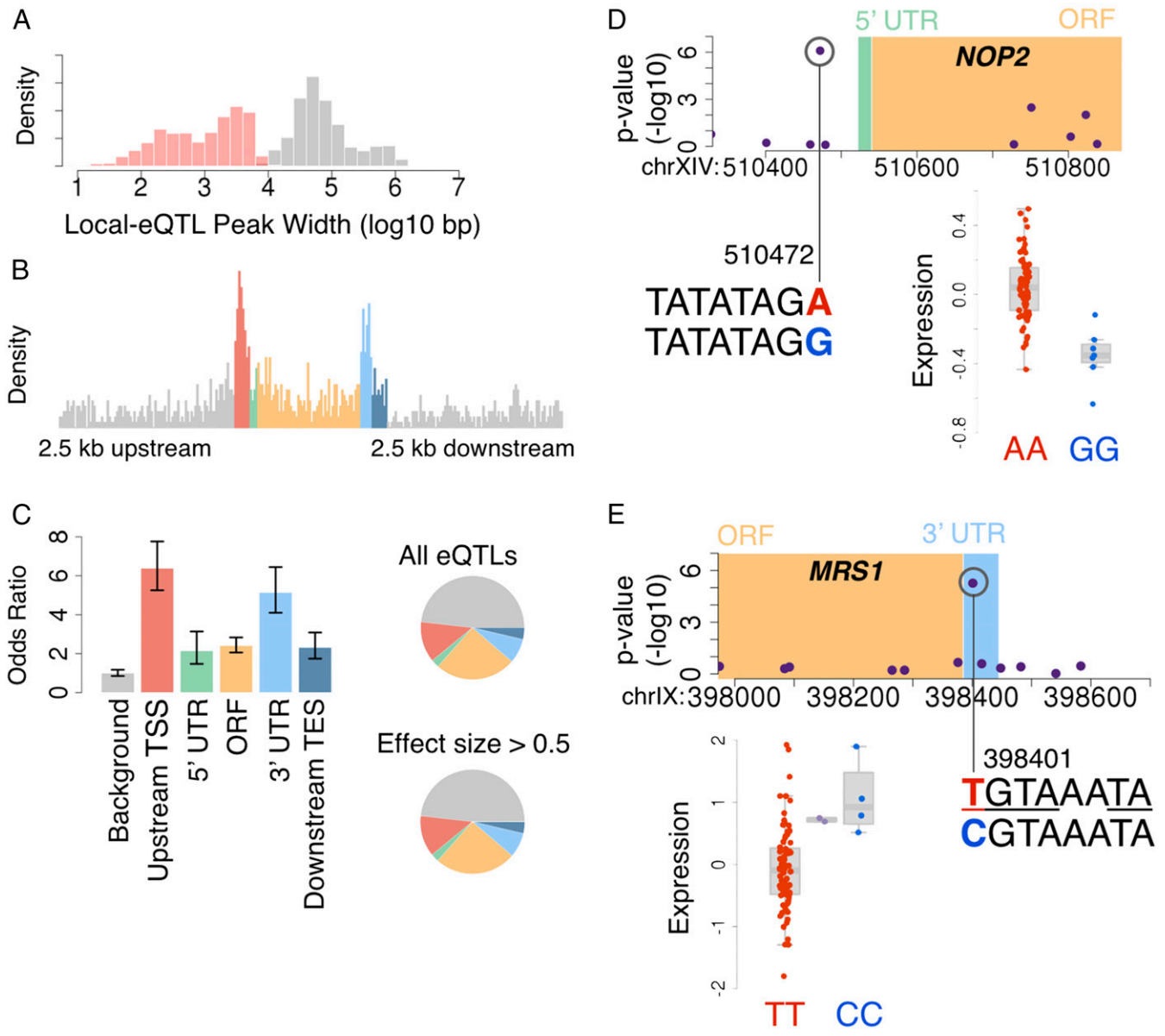

Fig. 2. High-resolution eQTL mapping. (*A*) Distribution of eQTL resolution width in the segregant analysis (gray) and this study (red). (*B*) Distribution of eQTLs relative to transcript components labeled in *C*. The upstream TSS and downstream TES refer to 1-kb windows. The length of each component has been scaled (set to the mean length of that component across all genes) so that all genes are displayed with the same length for each component. (*C*) Odds ratio of eQTL density of transcript component relative to expected. Pie chart of the proportion of transcript components for all eQTLs and eQTLs with large effect size. (*D* and *E*) eQTL for *NOP2* and *MRS1*. Purple points indicate the *P* value for each variant that was tested. The letters indicate the sequence of the region neighboring the variant with the most significant *P* value, and the colored letters indicate the alleles for that variant. The box plot depicts the normalized expression levels for isolates with each allele.

Taking advantage of this high resolution, we investigated the location of the strongest eQTL for each gene relative to several transcript components: the transcription start site (TSS), 5′UTR, ORF, 3′UTR, and transcription end site (TES). Previous analysis of 22 yeast strains did not detect enrichment of eQTLs in a particular component (28). We found the strongest enrichments upstream of the TSS (i.e., the promoter) and within the 3′UTR (Fig. 2*B*). To account for SNP density differences between components, we calculated the odds ratio of the enrichment compared with that expected by chance, which produced similar results (Fig. 2*C*). To evaluate whether variants in some locations might be more likely to result in strong effects on transcript levels, we examined the locations of eQTLs with large effect size ($\beta > 0.5$, 176 genes) (Fig. 2*C*). We found that the locations were similar to weaker eQTLs, suggesting that large-effect eQTLs are no more likely to occur in a particular location than weak eQTLs (*SI Appendix*, Fig. S12*B*).

An example of our high-resolution eQTL mapping is an eQTL for the gene *NOP2* (Fig. 2*D*). The strongest associating variant for *NOP2* was 50-bp upstream of the TSS, and the neighboring SNPs had far weaker associations, allowing us to define the likely causal SNP. Notably, this variant is found in a TATA-element as determined by a ChIP-exo study of preinitiation complex protein binding (29). Another illustrative example is an eQTL for the gene *MRS1*. Here, the strongest associating variant is within the 3′UTR, and the neighboring SNPs also have much weaker associations. This variant is located within an RNA-binding motif for PUF proteins: UGUA–UA (30). In addition, Puf3p binds to this transcript (31), and this variant is found four bases away from a Puf3p

binding site identified by PAR-Clip (global photoactivatable-ribonucleoside–enhanced cross-linking and immunopurification) (32).

Surprisingly, the 3′UTR enrichment is almost as strong as the enrichment upstream of the TSS (Fig. 2*C*). The 3′UTR is the site of many mRNA-protein interactions that affect RNA decay (33), thus we investigated whether there was any enrichment for particular binding sites among the 3′UTR eQTLs. We performed a de novo differential motif analysis in these regions, where we controlled for motifs that are enriched in 3′UTRs overall by using all 3′UTRs as the background (*SI Appendix*, *Text S4*). This analysis identified two significant motifs, which contained the RNA-binding motifs of Puf3p (UGUA) and Puf2p (UAAU) (31, 34) (*SI Appendix*, Fig. S13). Performing a similar de novo motif analysis on the promoter regions, we found strong enrichment of several motifs, although we do not yet know what functions these may have (*SI Appendix*, Fig. S14).

In addition to the enrichment in the 3′UTR and promoter regions, we also identified 310 eQTLs in the ORFs of their target genes, with increasing enrichment near the 5′ end. Regulatory variants in the ORF may have either a *cis*-acting mechanism (such as disruption of transcription factor binding), or a *trans*-acting effect (such as a nonsynonymous mutation affecting a gene's autoregulation, as seen for *AMN1*) (10). If such a self-regulating *trans*-mechanism was a common effect, we would expect an enrichment of nonsynonymous variants, since these are more likely to disrupt protein function than synonymous variants. Among all eQTLs within ORFs, we found no enrichment of nonsynonymous compared with synonymous variants (67.7% synonymous eQTL variants observed, compared with 66.9% expected by chance based on all variants in these genes), suggesting that most eQTLs within ORFs are unlikely to act *in trans*.

Species-Wide Selection Pressures on eQTLs. Gene expression has been shown to be predominantly under stabilizing selection in several species (35–37), although positive selection is also acting on the expression of many genes in yeast (13). To understand the evolutionary pressures on the eQTLs in this study, we evaluated the eQTLs for overall selection using two metrics: (*i*) the allele frequencies of the eQTLs and (*ii*) the presence and absence of eQTLs. Previous studies have found that eQTLs in humans and *Capsella grandiflora* are enriched for rare alleles, suggesting that eQTLs are under negative selection (38, 39). In yeast, a similar pattern of broad negative selection has been observed (40); however, because only three *S. cerevisiae* genomes were available at the time, rare alleles could not be confidently identified (40). With many more genomes and a wider sampling of gene-expression variation, we revisited this question with increased power and resolution.

To identify whether eQTLs are either depleted or enriched for rare alleles, a null distribution of expected minor allele frequencies (MAFs) is essential. To account for the statistical bias that eQTLs are more easily detected at high MAF, we generated an appropriate null-distribution of MAFs (39) (*SI Appendix*, *Text S5*). Comparing the MAF distribution of the randomized vs. real eQTLs, we found an enrichment of eQTLs among rare alleles (MAF < 0.15, $P < 10^{-3}$) (Fig. 3*A*), and a depletion of eQTLs among common alleles (MAF > 0.45, $P < 10^{-3}$) (Fig. 3*A*). These

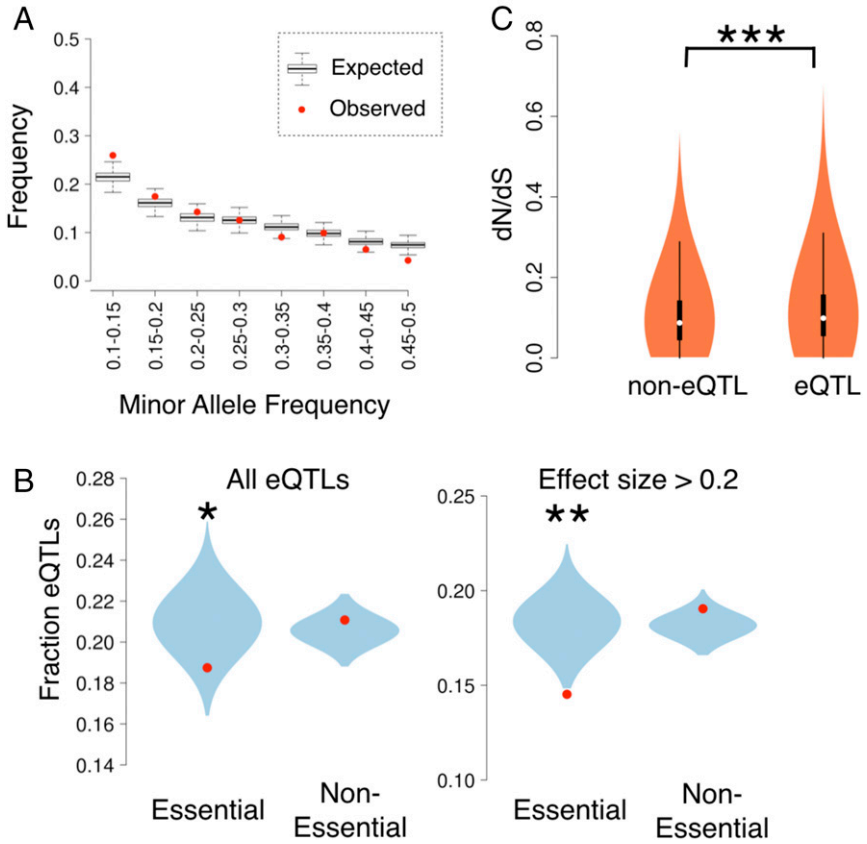

Fig. 3. Selection on eQTLs by changes in presence/absence and allele frequency. (*A*) The proportion of eQTLs observed within the specified MAF range is plotted in red. The boxplots represent the null distribution of expected minor allele frequencies. (*B*) The observed fraction of eQTLs among the genes is plotted in red. The blue violin plots depict the distribution of expected fraction of eQTLs from permutations that control for expression level. $^{**}P < 10^{-3}$ and $^{*}P < 0.05$. (*C*) The distribution of dN/dS values of genes as violin plots segregated by presence and absence of eQTLs. $^{***}P$ value $< 10^{-4}$.

analyses confirm that the main selective force on eQTLs in *S. cerevisiae* is negative selection.

We then hypothesized that the regulation of essential genes may be under greater negative selection than nonessential genes (41, 42). If so, mutations affecting expression of these genes would be quickly removed from the population, resulting in a dearth of detectable eQTLs (40). To test this, we compared the fractions of essential genes and nonessential genes among our eQTLs, while controlling for expression level (*SI Appendix*, *Text S5*). We observed a slightly lower number of eQTLs in essential genes (permutation $P = 0.036$) (Fig. 3*B*), which became more pronounced for stronger eQTLs (permutation $P = 4 \times 10^{-4}$) (Fig. 3*B*). eQTLs for essential genes also had a significantly smaller effect sizes compared with nonessential genes (median effect size 0.30 vs. 0.35, Mann–Whitney $P = 1.5 \times 10^{-5}$). Similarly, we found that genes with more constrained protein sequences [as measured by the dN/dS ratio (43)] also tend to have fewer eQTLs (permutation $P = 2 \times 10^{-4}$) (Fig. 3*C*), suggesting a congruence in negative selection on expression levels and protein sequences.

Molecular Differences Between Clinical and Nonclinical Isolates. Although our analysis above confirmed that eQTLs are generally under negative selection, previous work has found that yeast *cis*-regulation can also be the source of adaptations (18, 19). To explore this possibility, we next investigated whether any eQTLs have been subject to different evolutionary pressures between clinical and nonclinical isolates. *S. cerevisiae* is generally regarded as safe, but there have been case reports of invasive infection by *S. cerevisiae* across a wide range of patient types and locations (44, 45). These infections range from fungemia to endocarditis and are associated with conditions such as the placement of an intravenous catheter or an immune-compromised state. As a result, *S. cerevisiae* has been classified as an emerging opportunistic pathogen. This classification of a genetically tractable organism presents a unique opportunity for analyzing the evolutionary adaptations associated with pathogenicity.

We first sought to conduct a GWAS to identify genetic variants associated with the clinical niche. Before performing the GWAS, we investigated the statistical power to detect associations given the population structure. Although population structure correction is prudent in GWAS of any species, this is particularly important in *S. cerevisiae* because of widespread admixture between strains (27). Using simulations of the population structure specific to our strains, we found that we have power to detect associations across a range of effect sizes (*Methods* and *SI Appendix*, Fig. S15).

Our GWAS identified two SNPs reaching genome-wide significance (*Methods* and Fig. 4*A*) ($P < 1.1 \times 10^{-6}$, FDR < 0.05). One SNP (chromosome XII, position 830378) exhibited a particularly strong association ($P = 2.9 \times 10^{-12}$). The associated SNP is located 14-bp upstream of the ORF for *NIT3*, suggesting a *cis*-regulatory effect, and the lack of nearby associated variants (Fig. 4*B*) suggests that this SNP is likely to be the causal variant. The MAF is 0.41 in the clinical strains and 0.07 in the nonclinical strains. To test whether this variant could possibly be caused by sequencing error, we examined its read coverage and local linkage disequilibrium structure; both analyses confirmed this is a high-confidence variant (*SI Appendix*, *Text S6* and Figs. S16 and S17). This variant was not previously detected, perhaps due to the lower resolution or accuracy of tiling array-based genotyping (22). Investigation of the population variation of this SNP suggests that the variant likely

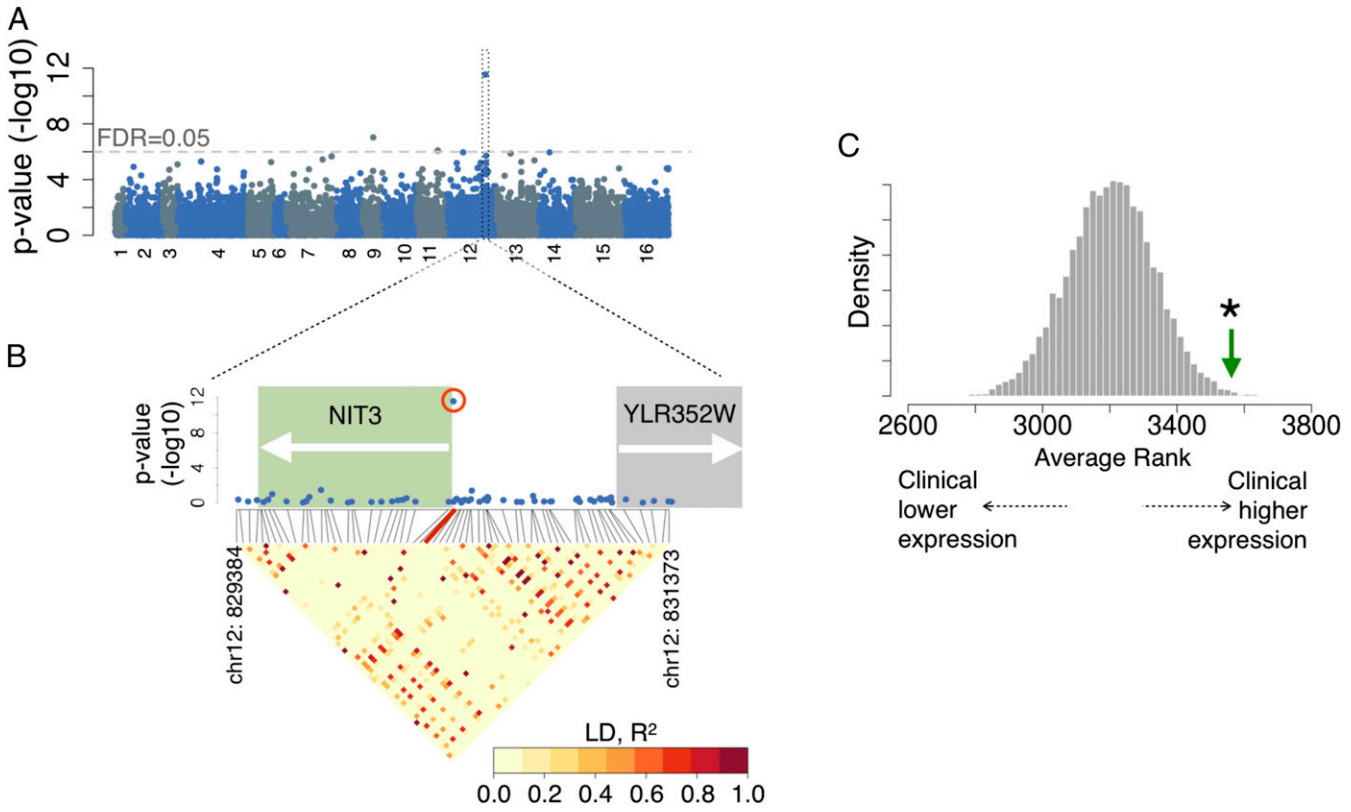

Fig. 4. GWAS and differential expression of biofilm genes between clinical and nonclinical strains. (*A*) Manhattan plot of clinical and nonclinical variants. (*B*) Higher detail of the region neighboring the strongest GWAS variant. The locations of genes are shown behind the Manhattan plot and the linkage disequilibrium is shown below. (*C*) Expression of biofilm suppressor gene set. The *x* axis is the average rank of differential expression across the gene set. The gene with the most significant lower clinical expression has a rank of zero. Gray histogram presents the distribution from null permutations of the average rank from gene sets of the same size. The green arrow indicates the observed average rank. *$P < 0.05$.

originated from a single mutation, as opposed to multiple independent mutations (*SI Appendix*, *Text S7*).

NIT3 is a member of the nitrilase superfamily, which encodes deaminating enzymes that play a role in biosynthesis of products such as biotin and auxin (46). The null mutant exhibits increased formation of biofilms (47), which are aggregate communities of adherent cells. Thus, we classified the gene as a biofilm suppressor. Biofilm formation has been associated with virulence in *Candida albicans* and other fungi (48, 49). In *S. cerevisiae*, biofilm formation is assessed by the ability to adhere to plastic, and biofilm formation is associated with decreased growth rate, increased drug resistance, and a "mat"-like appearance (50, 51).

We next investigated whether gene expression differs between clinical and nonclinical strains using DESeq2 (52) (*SI Appendix*, Fig. S19). At an FDR of 0.05, we identified 325 differentially expressed genes (Dataset S2). Because *NIT3* is a biofilm suppressor, we asked whether biofilm suppressors as a group exhibited differential expression between the clinical and nonclinical strains: 197 genes, including *NIT3*, were classified as biofilm suppressors because the deletion of the gene resulted in increased biofilm formation (47, 53). Comparing normalized expression levels between the clinical and nonclinical strains, we observed significantly increased expression of biofilm suppressors in the clinical strains (Mann–Whitney $P = 1.7 \times 10^{-5}$, median $\log_2$ fold-change 0.064), although *NIT3* was not differentially expressed. Because this observation could be driven by a small subset of genes, we also performed a more conservative test. For this test, we first ranked all genes based on their significance and directionality of differential expression between clinical vs. nonclinical strains. We then tested whether the average rank of biofilm suppressors was significantly different from the average rank of random sets of the same number of genes. Again, biofilm suppressors were significantly overexpressed in the clinical strains (permutation $P = 1.9 \times 10^{-3}$) (Fig. 4*C*). The increased expression of the biofilm suppressors in the clinical strains suggests that the clinical strains have an overall gene-expression profile consistent with disruption of biofilm formation.

Evidence for Lineage-Specific Selection on Biofilm Suppressors. We then proceeded to test whether this expression difference in biofilm suppressors was driven by natural selection acting on eQTLs. This question is important because although we found an overall gene-expression difference in biofilm suppressors, this is not by itself an indication of gene-expression adaptation. By chance, neutral drift of even a single variant (such as a mutation in a transcription factor) could single-handedly account for the differential expression of many genes. One approach to surmount this challenge is to identify an excess of independent regulatory variants that act in the same direction (e.g., up-regulation in clinical isolates), since this would not be expected under neutral evolution (54).

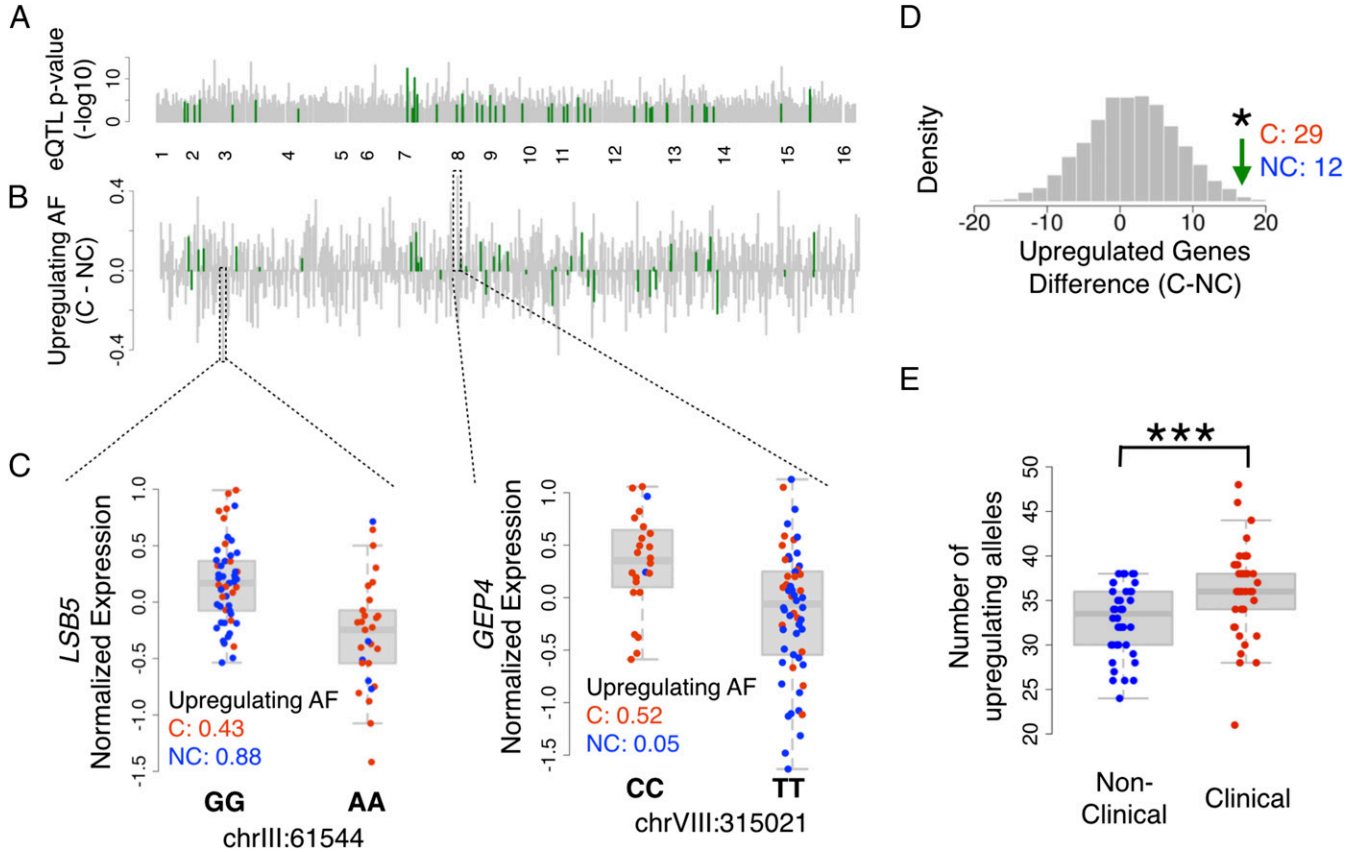

Fig. 5. Testing for adaptation in biofilm suppressors. (*A*) Locations of *cis*-eQTLs across the genome (FDR < 0.05). The biofilm suppressors that have significant *cis*-eQTLs are highlighted in green. (*B*) Difference in up-regulating allele frequency between the clinical and nonclinical strains for each of the *cis*-eQTLs. (*C*) Examples of *cis*-eQTLs with the largest positive and negative difference in up-regulating allele frequencies. Each point represents a strain, with the color indicating whether the strain is clinical (red) or nonclinical (blue). (*D*) Permutation test to assess the significance of the number of positive up-regulating allele frequencies in the clinical strains. The *x* axis indicates the difference in the number of genes that have higher up-regulating allele frequencies in the clinical strains and the number of genes that have higher up-regulating allele frequencies in the nonclinical strains. The green arrow marks the difference found in biofilm suppressors, and the histogram indicates the difference found across each of 1,000 permutations. (*E*) The number of up-regulating alleles for each strain segregated by clinical and nonclinical origin. *$P < 0.05$; ***$P < 10^{-3}$.

For each eQTL, we assessed whether the clinical strains exhibited a general up- or down-regulation of the gene by calculating the difference in the up-regulating allele frequency among the clinical strains and the nonclinical strains (Fig. 5 *A* and *B*). An overall positive value indicates that the clinical strains had a higher up-regulating allele frequency, suggesting that the local genetic control of that gene is shifted toward up-regulation in the clinical strains. The eQTLs with the largest positive and negative allele frequency differences are shown as examples (Fig. 5*C*). Interestingly, both genes are involved in endocytosis, which is a function that has been previously identified to be under lineage-specific selection in clinical strains (17).

Using these directionalities, we next investigated the overall directionality of the local eQTLs in biofilm suppressors. Among the biofilm suppressor genes, we identified 29 local eQTLs with higher up-regulating frequency in the clinical strains, and 12 local eQTLs with higher up-regulating frequency in the nonclinical strains. To assess the significance of this difference, we repeated the analysis with 10,000 random sets of 41 local eQTLs. This permutation procedure accounts for population structure and demographic confounders because the difference within the biofilm suppressors is compared with the rest of the genome (method adapted from ref. 7). Using this test, we found significantly more clinical up-regulated biofilm suppressors than expected (permutation $P = 0.019$) (Fig. 5*D*). We confirmed this result using multiple different cut-offs to define eQTLs [$P = 0.016$ with empirical eQTL $P < 0.1$ and $P = 0.029$ with empirical eQTL $P < 0.2$, which is the least stringent cut-off at which there is significant concordance in directionality with the Skelly et al. (28) dataset] (*SI Appendix*, Fig. S10). The occurrence of this many independent regulatory variants shifting in the same direction (up-regulation in clinical isolates) is unlikely to occur under neutral evolution. Analogous to the sign test in allele-specific expression analyses, this indicates that the biofilm suppressors are likely under lineage-specific selection. Since human infection is almost certainly not the ancestral niche of *S. cerevisiae*, we can additionally infer that the change is most likely due to up-regulation in response to novel selection pressures on the clinical strains, rather than down-regulation in nonclinical strains.

We next assessed whether the allele frequency difference was occurring in a small sample of clinical strains, or if this was a general trend across all clinical strains. To investigate this distribution, we counted the number of up-regulating biofilm eQTL alleles for each strain (Fig. 5*E*). We observed a general shift in distribution, suggesting that the overall up-regulation occurred as a trend across the majority of clinical strains (Mann–Whitney $P = 8.1 \times 10^{-4}$). Concordant with the differential expression analysis, these results suggest that the clinical strains have undergone adaptation to increase expression of the biofilm suppressors.

Discussion

In this work, we identified local eQTLs regulating ∼20% of the genes in *S. cerevisiae*. Although previous yeast eQTL studies have revealed rich information on the genetics of gene expression (9, 11, 28, 55), mapping with 85 diverse strains has the advantages of greater genetic diversity and less linkage disequilibrium between nearby variants. These qualities facilitated fine-mapping of causal eQTL variants and also allowed us to explore the species-wide evolution of eQTLs.

Assessing all of the eQTLs, we found that their 1-LOD support intervals are ∼50-fold smaller than previous yeast eQTLs—with 33% mapped to single-variant resolution—and that they are enriched near the TSSs and TESs of genes. Previous studies in several species have found eQTL enrichment near TSSs and TESs (57–61), as well as higher genetic variation in 3′ UTRs of genes with allele-specific expression (10); however, to our knowledge our eQTLs represent a unique example of eQTLs implicating a specific molecular mechanism within 3′ UTRs. We found that the 3′ UTR eQTLs are enriched for RNA-binding protein motifs, suggesting that a common mechanism for eQTLs is posttranscriptional. We did not find any evidence for eQTLs within ORFs to act *in trans*, since there was no enrichment for nonsynonymous variants that would be much more likely to mediate *trans*-acting effects.

We used these high-resolution eQTLs to study the evolution of gene expression across our 85 strains. Previous studies in several species have shown that gene expression is predominantly under stabilizing selection, for example, in mutation accumulation experiments where organisms are grown for many generations with minimal selection (35, 36). Indeed, we confirmed that eQTLs are generally under negative selection (38–40), and that eQTLs are depleted among essential genes and evolutionarily constrained genes (40).

Despite the predominance of stabilizing selection, there is growing evidence that positive selection is also a widespread force acting on regulatory variation in a wide range of species (7, 8, 54). To search for gene-expression adaptation in these strains, we tested for lineage-specific selection on eQTLs between the clinical and nonclinical strains. Our GWAS and differential expression analysis independently implicated biofilm suppressors as a differentiating gene set. Consistent with this, we found that allele frequency shifts of eQTLs have led to increased expression of biofilm suppressor genes in clinical strains. Although the allele frequency shifts were generally small, by testing across a large number of genes, we were able to identify a significant pattern that is not consistent with neutral evolution (54).

Biofilm regulation has been previously studied in other medically relevant fungi (48), and has a particularly rich literature in *C. albicans* (49). Although *Candida* and *Saccharomyces* are evolutionarily distant, elements of the biofilm formation pathways are conserved between them (62). Notably, we found evidence for increased biofilm suppression in the clinical strains, which is contrary to the standard impression of biofilm activation in pathogenic microbes. We hypothesize several possible explanations: (*i*) biofilm dispersal is a key component of the biofilm virulence effect (49, 63, 64), and the increased expression of biofilm suppressors may promote this dispersal; or (*ii*) the classification of biofilm genes as general suppressors or activators may be too simplistic (e.g., ignoring condition-dependent effects) and thus, our results could reflect adaptation in a specific but uncharacterized component of the biofilm process. Furthermore, these genes are likely to have a variety of functions unrelated to biofilms, and thus the effects that we observed may even represent selection acting on a different phenotype. This hypothesis would also be consistent with the lack of signal observed when the sign test is performed on biofilm-promoters ($P = 0.68$). Future experiments will be essential in determining whether these eQTLs affect pathogenicity, or some other aspect of adaptation to human hosts.

In addition to further investigating the role of biofilm formation, future studies may reveal much more about these strains. For example, recent studies have found a large number of ploidy and copy-number variations between strains of *S. cerevisiae* (65, 66). In this study, although the absence and presence of genes were assessed, we did not estimate copy number (*SI Appendix*, *Text S2*). Instead, by measuring gene expression, we were able to directly measure the amount of mRNA and thus measure the downstream effects of whatever copy-number variation exists. The difference, however, between gene-expression regulation by eQTL versus copy-number variation remains to be investigated. Another area of future study is the measurement of gene expression in multiple conditions, since eQTLs can be environment-specific (11). In addition, work in *C. albicans* has found that gene regulatory relationships differ between infection environments and laboratory media (67). Such environment-specific expression might explain the lack of differential expression seen in *NIT3* [although this variant might also act at another level of regulation, such as

translation initiation, another major source of divergence in yeast gene regulation (68, 69)].

The strength of *S. cerevisiae* as a model organism is based upon its easy manipulation, well-scrutinized genes, and a rich landscape of strains from diverse geographical and ecological niches. In this study, we contribute to this compendium of knowledge with genome-wide gene expression across many strains and a high-resolution map of genetic regulation of gene expression. We have shown that gene expression across these strains is not only constrained, but also under pressure to adapt across the varied strain histories and phenotypes. We hope for this resource to serve as the launching point for future studies on the mechanisms and consequences of gene-expression variation.

Methods

Strain Selection and DNA Sequencing. Eighty-five isolates of *S. cerevisiae* were obtained from the Pfaff Yeast Culture Collection (*SI Appendix*, Table S1). Locations of isolation were obtained from Muller et al. (22). Isolates were grown overnight in YPD liquid culture at 30 °C, and DNA extraction was performed using the MasterPure Yeast DNA purification kit. Pooled, barcoded libraries were prepared using the Nextera Sample Preparation Kit. Sequencing was performed using Illumina Hi-Seq 2000 with paired-end 101-bp reads. Reads were then dynamically trimmed and length-sorted using SolexaQA default parameters (70). Mapping was performed using STAMPY with default parameters and substitution rate 0.005 (71). Reads were then sorted, and duplicates were removed using Samtools (72). Variant calling was performed using freeBayes (default parameters, $T = 0.005$) using all BAM files together (73). To compare the percent concordance of the allele calls in the strains that were sequenced in both this study and Strope et al. (23), we also performed mapping and variant calling on the Strope dataset using the same parameters. For downstream analysis using GEMMA, which requires no missing variants, we removed SNPs with greater than 10% missing data and imputed the rest of the missing variants using MACH (–mle–rounds 100–states 200) (74).

RNA Sequencing. Strains were grown to log-phase growth in YPD at 30 °C, and total RNA was extracted using the MasterPure Kit. Sequencing libraries were constructed using the Illumina TruSeq kit. Thirty-six–base pair single-end reads were sequenced using the Illumina HiSEq. 2000 across eight lanes with a randomized lane assignment to prevent batch effects. The reads were dynamically trimmed using DynamicTrim and LengthSort of the SolexaQA package with default parameters (70). Reads were then mapped using STAR v2.4.1 to the S288c reference genome with all SNPs from the DNA-sequencing analysis masked (75). The genome index was generated with flag "–genomeSAindexNbases 11" and mapping was performed with default parameters. STAR mitigates mapping bias by allowing for mismatches. The lack of genome-wide mapping bias was confirmed by a lack of significant correlation between percent reads mapping and the genomic distance from the S288c reference genome (*SI Appendix*, Fig. S4). Principal components of the expression data were calculated using the standardized reads per kilobase and million mapped reads (RPKM) expression, excluding genes with a sum of RPKM across all strains <1. In addition, we assessed whether genes were missing using the DNA sequencing data, and these genes were marked as missing for subsequent eQTL mapping (*SI Appendix*, Text S2).

eQTL Mapping. RPKM values were quantile-normalized across strains and the resulting values were fit to a standard normal for each gene. We then used PEER to discover hidden covariates ($k = 15$), and these covariates were regressed out (76). The removal of hidden covariates improves power in *cis*-eQTL mapping and also can remove unwanted batch effects (76–78). We mapped eQTLs using variants with MAF greater than 0.1 using GEMMA (26). Because GEMMA does not allow missing variants, we imputed the variants using MACH 1.0 (74), removing any variants with more than 10% missing. The relatedness matrix for association testing was calculated using all variants after linkage disequilibrium pruning ($r^2 > 0.8$, sliding window) with the centered relatedness approach, as described in the GEMMA manual. To assess significance, we performed 1,000–10,000 permutations, comparing for each gene, the best-associating P value from each permutation with the best-associating P value from the nonpermuted associations. The permutations provided empirical P values, from which we assessed FDRs using the Benjamini–Hochberg method (79). We compared results from these permutations performed on each transcript individually with that of permutations performed preserving the relationship across transcripts, which revealed similar empirical P values (*SI Appendix*, Fig. S11). There was a median of 12 variants within the 1-LOD support intervals (including 462 with exactly one variant), and a median of 34 variants within the 2-LOD support intervals.

To localize the eQTLs relative to genic components, we used annotations of ORFs from *Saccharomyces* Genome Database (80). In cases where multiple eQTLs for the same gene were tied for most significant, the midpoint between the minimum and maximum position was used to represent the eQTL location. Transcript boundaries of the isoform with the highest number of supporting reads were used (81). Synonymous vs. nonsynonymous variants were identified with SNPeff (82).

Genome-Wide Association Testing. Genome-wide association testing was performed using variants with no missing genotypes, MAF > 0.05, and linkage disequilibrium pruned by PLINK ($r^2 > 0.8$, sliding window) (83, 84). To identify the most-effective method for mapping between genotypes and traits, we performed a power analysis using three methods (*SI Appendix*, Text S8). We found that GEMMA performed best at controlling for false-positives due to population structure while maintaining power. We thus performed all subsequent association testing with GEMMA. The relatedness matrix was calculated using all tested variants with the centered relatedness approach, as described in the GEMMA manual. Calculating significance after GWAS used a simpler approach than eQTL mapping because all variants were tested together rather than using a gene-by-gene approach. The cut-off for FDR was identified by 10^5 permutations of a randomized phenotype, identifying 1.02×10^{-6} as the cut-off for FDR < 0.05.

Data. All RNA-Seq and DNA-Seq data are deposited in Bioproject PRJNA342356 in the National Center for Biotechnology Information Sequence Read Archive.

ACKNOWLEDGMENTS. We thank members of the H.B.F. laboratory for helpful discussions, and K. Boundy-Mills for providing yeast strains. This work was supported by NIH Grant 2R01GM097171-05A1.

1. Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16:197–212.
2. Musunuru K, et al. (2010) From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* 466:714–719.
3. Claussnitzer M, et al. (2015) FTO obesity variant circuitry and adipocyte browning in humans. *N Engl J Med* 373:895–907.
4. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. *Genome Res* 22: 1748–1759.
5. Gusev A, et al. (2016) Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 48:245–252.
6. Zhu Z, et al. (2016) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* 48:481–487.
7. Fraser HB (2013) Gene expression drives local adaptation in humans. *Genome Res* 23: 1089–1096.
8. Enard D, Messer PW, Petrov DA (2014) Genome-wide signals of positive selection in human evolution. *Genome Res* 24:885–895.
9. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–755.
10. Ronald J, Brem RB, Whittle J, Kruglyak L (2005) Local regulatory variation in *Saccharomyces cerevisiae*. *PLoS Genet* 1:e25.
11. Smith EN, Kruglyak L (2008) Gene-environment interaction in yeast gene expression. *PLoS Biol* 6:e83.
12. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436:701–703.
13. Fraser HB, Moses AM, Schadt EE (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci USA* 107:2977–2982.
14. Tirosh I, Reikhav S, Levy AA, Barkai N (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* 324:659–662.
15. Martin HC, Roop JI, Schraiber JG, Hsu TY, Brem RB (2012) Evolution of a membrane protein regulon in *Saccharomyces*. *Mol Biol Evol* 29:1747–1756.
16. Roop JI, Chang KC, Brem RB (2016) Polygenic evolution of a sugar specialization trade-off in yeast. *Nature* 530:336–339.
17. Fraser HB, et al. (2012) Polygenic *cis*-regulatory adaptation in the evolution of yeast pathogenicity. *Genome Res* 22:1930–1939.
18. Chang J, et al. (2013) The molecular mechanism of a *cis*-regulatory adaptation in yeast. *PLoS Genet* 9:e1003813.
19. Naranjo S, et al. (2015) Dissecting the genetic basis of a complex *cis*-regulatory adaptation. *PLoS Genet* 11:e1005751.
20. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322:881–888.
21. Flint J, Eskin E (2012) Genome-wide association studies in mice. *Nat Rev Genet* 13: 807–817.
22. Muller LA, Lucas JE, Georgianna DR, McCusker JH (2011) Genome-wide association analysis of clinical vs. nonclinical origin provides insights into *Saccharomyces cerevisiae* pathogenesis. *Mol Ecol* 20:4085–4097.

23. Strope PK, et al. (2015) The 100-genomes strains, an *S. cerevisiae* resource that illuminates its natural phenotypic and genotypic variation and emergence as an opportunistic pathogen. *Genome Res* 25:762–774.
24. Liti G, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458: 337–341.
25. Schacherer J, Shapiro JA, Ruderfer DM, Kruglyak L (2009) Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* 458:342–345.
26. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44:821–824.
27. Connelly CF, Akey JM (2012) On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*. *Genetics* 191:1345–1353.
28. Skelly DA, et al. (2013) Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. *Genome Res* 23:1496–1504.
29. Rhee HS, Pugh BF (2012) Genome-wide structure and organization of eukaryotic preinitiation complexes. *Nature* 483:295–301.
30. Hogan GJ, Brown PO, Herschlag D (2015) Evolutionary conservation and diversification of Puf RNA binding proteins and their mRNA targets. *PLoS Biol* 13: e1002307.
31. Gerber AP, Herschlag D, Brown PO (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* 2:E79.
32. Freeberg MA, et al. (2013) Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*. *Genome Biol* 14:R13.
33. Pai AA, et al. (2012) The contribution of RNA decay quantitative trait loci to interindividual variation in steady-state gene expression levels. *PLoS Genet* 8:e1003000.
34. Porter DF, Koh YY, VanVeller B, Raines RT, Wickens M (2015) Target selection by natural and redesigned PUF proteins. *Proc Natl Acad Sci USA* 112:15868–15873.
35. Denver DR, et al. (2005) The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. *Nat Genet* 37:544–548.
36. Rifkin SA, Houle D, Kim J, White KP (2005) A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* 438:220–223.
37. Landry CR, Lemos B, Rifkin SA, Dickinson WJ, Hartl DL (2007) Genetic properties influencing the evolvability of gene expression. *Science* 317:118–121.
38. Battle A, et al. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res* 24:14–24.
39. Josephs EB, Lee YW, Stinchcombe JR, Wright SI (2015) Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc Natl Acad Sci USA* 112:15390–15395.
40. Ronald J, Akey JM (2007) The evolution of gene expression QTL in *Saccharomyces cerevisiae*. *PLoS One* 2:e678.
41. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. *PLoS Biol* 2:e137.
42. Tirosh I, Barkai N (2008) Evolution of gene sequence and gene expression are not correlated in yeast. *Trends Genet* 24:109–113.
43. Connallon T, Knowles LL (2007) Recombination rate and protein evolution in yeast. *BMC Evol Biol* 7:235.
44. Enache-Angoulvant A, Hennequin C (2005) Invasive *Saccharomyces* infection: A comprehensive review. *Clin Infect Dis* 41:1559–1568.
45. Muñoz P, et al. (2005) *Saccharomyces cerevisiae* fungemia: An emerging infectious disease. *Clin Infect Dis* 40:1625–1634.
46. Pace HC, Brenner C (2001) The nitrilase superfamily: Classification, structure and function. *Genome Biol* 2:REVIEWS0001.
47. Vandenbosch D, et al. (2013) Genomewide screening for genes involved in biofilm formation and miconazole susceptibility in *Saccharomyces cerevisiae*. *FEMS Yeast Res* 13:720–730.
48. Fanning S, Mitchell AP (2012) Fungal biofilms. *PLoS Pathog* 8:e1002585.
49. Nobile CJ, Johnson AD (2015) *Candida albicans* biofilms and human disease. *Annu Rev Microbiol* 69:71–92.
50. Reynolds TB, Fink GR (2001) Bakers' yeast, a model for fungal biofilm formation. *Science* 291:878–881.
51. Bojsen R, Regenberg B, Folkesson A (2014) *Saccharomyces cerevisiae* biofilm tolerance towards systemic antifungals depends on growth phase. *BMC Microbiol* 14:305.
52. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15:550.
53. Scherz K, et al. (2014) Genetic basis for *Saccharomyces cerevisiae* biofilm in liquid medium. *G3 (Bethesda)* 4:1671–1680.
54. Fraser HB (2011) Genome-wide approaches to the study of adaptive gene expression evolution: Systematic studies of evolutionary adaptations involving gene expression will allow many fundamental questions in evolutionary biology to be addressed. *Bioessays* 33:469–477.
55. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci USA* 102:1572–1577.
56. Gagneur J, et al. (2013) Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype. *PLoS Genet* 9:e1003803.
57. Veyrieras JB, et al. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet* 4:e1000214.
58. Stranger BE, et al. (2012) Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet* 8:e1002639.
59. Veyrieras JB, et al. (2012) Exon-specific QTLs skew the inferred distribution of expression QTLs detected using gene expression array data. *PLoS One* 7:e30629.
60. Huang W, et al. (2015) Genetic basis of transcriptome diversity in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 112:E6010–E6019.
61. Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y (2015) The genetic architecture of gene expression levels in wild baboons. *eLife* 4.
62. Nobile CJ, et al. (2012) A recently evolved transcriptional network controls biofilm development in *Candida albicans*. *Cell* 148:126–138.
63. Uppuluri P, et al. (2010) Dispersion as an important step in the *Candida albicans* biofilm developmental cycle. *PLoS Pathog* 6:e1000828.
64. Uppuluri P, Lopez-Ribot JL (2016) Go forth and colonize: Dispersal from clinically important microbial biofilms. *PLoS Pathog* 12:e1005397.
65. Dunn B, Richter C, Kvitek DJ, Pugh T, Sherlock G (2012) Analysis of the *Saccharomyces cerevisiae* pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Res* 22:908–924.
66. Bergström A, et al. (2014) A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol* 31:872–888.
67. Xu W, et al. (2015) Activation and alliance of regulatory pathways in *C. albicans* during mammalian infection. *PLoS Biol* 13:e1002076.
68. Artieri CG, Fraser HB (2014) Evolution at two levels of gene expression in yeast. *Genome Res* 24:411–421.
69. McManus CJ, May GE, Spealman P, Shteyman A (2014) Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* 24: 422–430.
70. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* 11:485.
71. Lunter G, Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21:936–939.
72. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079.
73. Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907v2.
74. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34: 816–834.
75. Dobin A, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21.
76. Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput Biol* 6:e1000770.
77. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3:1724–1735.
78. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118–127.
79. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc B* 57:289–300.
80. Cherry JM, et al. (2012) *Saccharomyces* Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res* 40:D700–D705.
81. Pelechano V, Wei W, Steinmetz LM (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 497:127–131.
82. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6:80–92.
83. Purcell S, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
84. Chang CC, et al. (2015) Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* 4:7.